Published in "International Journal of Modern Physics C 21(6): 813-824 , 2010"
which should be cited to refer to this work.

SIMILARITY-BASED CLASSIFICATION IN PARTIALLY LABELED NETWORKS

QIAN-MING ZHANG* and MING-SHENG SHANG†

Web Sciences Center, School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu 610054, P. R. China
** air_spire@163.com*
† msshang@uestc.edu.cn

LINYUAN LÜ‡

Department of Physics, University of Fribourg
Chemin du Musée 3, Fribourg CH-1700, Switzerland
linyuan.lue@unifr.ch

Two main difficulties in the problem of classification in partially labeled networks are the sparsity of the known labeled nodes and inconsistency of label information. To address these two difficulties, we propose a similarity-based method, where the basic assumption is that two nodes are more likely to be categorized into the same class if they are more similar. In this paper, we introduce ten similarity indices defined based on the network structure. Empirical results on the co-purchase network of political books show that the similarity-based method can, to some extent, overcome these two difficulties and give higher accurate classification than the relational neighbors method, especially when the labeled nodes are sparse. Furthermore, we find that when the information of known labeled nodes is sufficient, the indices considering only local information can perform as good as those global indices while having much lower computational complexity.

Keywords: Complex networks; similarity index; classification; labeled networks.

PACS Nos.: 89.20.Ff, 89.75.Hc, 89.65.-s.

1. Introduction

Recently, the problem of within-network classification in partially labeled networks has attracted much attention. Given a network with partial nodes being labeled, the problem is to predict the labels of these unlabeled nodes based on the known labels and the network structure. Many algorithms have been proposed. These methods

‡Corresponding author.

can be widely applied to many fields, such as the hypertext categorization,[1,2] distinguishing the fraud and legit users in cell phone network,[3] detecting whether an email is for a certain task[4] and predicting the disease-related genes.[5] Generally speaking, the known methods can be classified into two groups. One is collective classification, which refers to the combined classification by using three types of correlations:

(i) between the node's label and its attributes,
(ii) between node's label and its neighbor's attributes,
(iii) between node's label and its neighbor's label (see a brief introduction in Ref. 6).

One remarkable advantage of this method is its high ability to learn the dependency structure, such as positive or negative correlation (i.e. consistency or inconsistency). However, when the labeled nodes are sparse, this method is difficult to give accurate classification. The sparse problem can be solved by another group of methods, named semi-supervised learning, which make use of both labeled and unlabeled data for training (see Ref. 7 for more information). The latent assumption of this method is the consistency with the label information, namely the nearby nodes tend to have the same label. Therefore when this assumption does not hold the performance of this method will be largely degraded. Brian *et al.* proposed a method by adding ghost edges between every pair of labeled and unlabeled nodes to the target network, which enable the flow of information from the labeled nodes to the unlabeled nodes.[3] They assigned a weight to each ghost edge based on the score of the two endpoints obtained by the *Even-step random walk with restart* (*Even-step RWR*) algorithm. The experimental results on real-world data showed that their method can, to some extent, solve the sparse problem and negative correlation problem (i.e. inconsistency), and perform well while the existing approaches, such as collective classification and semi-supervised learning, will fail. In this paper, we compare the performances of *Even-step RWR* index with other nine similarity indices which have been widely used in link prediction problem.[8–10] These include five local indices, namely the *Common Neighbors*,[11] *Jaccard coefficient*,[12] *Sørensen index*,[13] *Adamic-Adar index*[14] and *Resource Allocation index*,[9] and four global indices, namely *Katz index*,[15] *Average Commute Time*,[16] *cosine based on the Pseudoinverse of the Laplacian matrix* ($\cos^+$) and *Random Walk with Restart* (*RWR*).[17] In addition, we also consider a simple *Relational Neighbors* algorithm, which claims that an unlabeled node tends to have the same label with its neighbors.[18] Empirical results on the co-purchase network of political books show that the similarity-based methods perform better overall than the *Relational Neighbors* algorithm. Especially when the labeled nodes are sparse, the improvement is prominent. Furthermore, when the data is dense, the best-performing local index can predict as good as the global indices and sometimes even better. However when the data is sparse, the best-performing global index will predict better than the best-performing local index.

The rest of this paper is organized as follows. In Sec. 2 we introduce ten similarity indices, including five indices based on local information and others based on global information. Section 3 describes the metric to evaluate the algorithm's accuracy. Section 4 shows the experimental results of the ten indices on the co-purchase network of political books. Finally, we conclude this paper in Sec. 5.

2. Similarity Indices

We consider five local similarity indices as well as five global ones. All are defined based on the network structure. A short introduction of each index is shown as:

(1) *Common Neighbors* (*CN*) — For a node x, let $\Gamma(x)$ denote the set of neighbors of x. By common sense, two nodes, x and y, are more similar if they have more common neighbors. The simplest measure of this neighborhood overlap is the directed count, namely

$$s_{xy}^{CN} = |\Gamma(x) \cap \Gamma(y)|\,, \tag{1}$$

where $|Q|$ is the cardinality of the set Q. It is obvious that $s_{xy}^{CN} = (A^2)_{xy}$, where A is the adjacency matrix, in which $A_{xy} = 1$ if x and y are directly connected and $A_{xy} = 0$ otherwise. Note that, $(A^2)_{xy}$ is also the number of different paths with length 2 connecting x and y.

(2) *Jaccard index*[12] — This index was proposed by Jaccard over a hundred years ago, and is defined as

$$s_{xy}^{\text{Jaccard}} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}\,. \tag{2}$$

(3) *Sørensen index*[13] — This index is used mainly for ecological community data, and is defined as

$$s_{xy}^{\text{Sørensen}} = \frac{2|\Gamma(x) \cap \Gamma(y)|}{k(x) + k(y)}\,. \tag{3}$$

(4) *Adamic-Adar index* (*AA*)[14] — This index refines the simple counting of common neighbors by assigning the less-connected neighbors more weight, and is defined as:

$$s_{xy}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k(z)}\,. \tag{4}$$

(5) *Resource Allocation* (*RA*)[9] — Consider a pair of nodes, x and y, which are not directly connected. The node x can send some resource to y, with their common neighbors playing the role of transmitters. In the simplest case, we assume that each transmitter has a unit of resource, and will equally distribute it between all its neighbors. The similarity between x and y can be defined as the amount of resource y received, which is:

$$s_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k(z)}\,. \tag{5}$$

Clearly, this measure is symmetric, namely $s_{xy} = s_{yx}$. Note that, although resulting from different motivations, the AA index and RA index have very similar forms. Indeed, they both depress the contribution of the high-degree common neighbors in different ways. AA index takes the $\log k(z)$ form while RA index takes the linear form. The difference is insignificant when the degree, k, is small, while it is great when k is large. Therefore, RA index punishes the high-degree common neighbors heavily.

(6) *Katz index*[15] — This measure is based on the ensemble of all paths, which directly sums over the collection of paths and exponentially damped by length to give the short paths more weights. The mathematical expression reads

$$s_{xy}^{\mathrm{Katz}} = \sum_{l=1}^{\infty} \beta^l \cdot |\mathrm{paths}_{xy}^{\langle l \rangle}| = \beta A + \beta^2 A^2 + \beta^3 A^3 + \cdots , \tag{6}$$

where $\mathrm{paths}_{xy}^{\langle l \rangle}$ is the set of all paths with length l connecting x and y, and β is a free parameter controlling the weights of the paths. Obviously, a very small β yields a measure close to CN, because the long paths contribute very little. The similarity matrix S, whose elements are s_{xy}^{Katz}, can be written as $(I - \beta A)^{-1} - I$, where I is the identity matrix. Note that β must be lower than the reciprocal of the maximum of the eigenvalues of matrix A to ensure the convergence.

(7) *Average Commute Time* (ACT)[16] — Denoting by $m(x, y)$ the average number of steps required by a random walker starting from node x to reach node y, the average commute time between x and y is $n(x, y) = m(x, y) + m(y, x)$, which can be computed in terms of the Pseudoinverse of the Laplacian matrix L^+ (see footnote[a]), as:

$$n(x, y) = E(l_{xx}^+ + l_{yy}^+ - 2l_{xy}^+) , \tag{7}$$

where E is the number of links in the network, l_{xy}^+ denotes the corresponding entry in L^+. Assuming two nodes are considered to be more similar if they have a small average commute time, then the similarity between the nodes x and y can be defined as the reciprocal of $n(x, y)$, namely (the constant factor E is removed).

$$s_{xy}^{ACT} = \frac{1}{l_{xx}^+ + l_{yy}^+ - 2l_{xy}^+} . \tag{8}$$

(8) *Cosine based on* L^+($\cos^+$)[16] — This index is an inner-product-based measure, which is defined as the **cosine** of node vectors, namely

$$s_{xy}^{\cos^+} = \cos(x, y)^+ = \frac{l_{xy}^+}{\sqrt{l_{xx}^+ \cdot l_{yy}^+}} . \tag{9}$$

(9) *Random walk with restart* (RWR)[17] — This index is a direct application of the PageRank algorithm. Consider a random walker starting from node x, who will iteratively move to a random neighbor with probability c and return to node x with

[a] $L = D - A$, where D is the degree matrix with $D_{ij} = \delta_{ij} \cdot k_i$.

probability $1-c$. Denote by q_{xy} the probability this random walker locates at node y in the steady state, then we have

$$\mathbf{q}_x = cP^T\mathbf{q}_x + (1-c)\mathbf{e}_x\,, \tag{10}$$

where $\mathbf{e}_x$ is an $N \times 1$ vector with the xth element equal to 1 and others all equal to 0, and $P^T = AD^{-1}$ where $D_{ij} = \delta_{ij}k_i$. The solution is straightforward, as

$$\mathbf{q}_x = (1-c)(I - cP^T)^{-1}\mathbf{e}_x\,. \tag{11}$$

Then the similarity between node x and y equals $s_{xy} = q_{xy} + q_{yx}$.

(10) *Even-step RWR*[3] — To avoid the immediate neighbors, we only consider the even-length paths. Mathematically, we should replace the transition matrix with $M = (P^T)^2$.

For comparison, we compare the above-mentioned ten indices with the simplest method, say *Relational Neighbors* (*RN*).[18] Given an unlabeled node u, the probability that its label is l_i equals

$$p(l_i|u) = \frac{|V'|}{|V''|}\,, \tag{12}$$

where V' is the set constituted by u's neighbors whose label is l_i, and V'' is the set of u's neighbors being labeled.

3. Method

Consider an unweighted undirected network of both labeled and unlabeled nodes: $G(V, E, L)$, where V is the set of nodes, E is the set of links and $L = \{l_1, l_2, \ldots, l_m\}$ is the set of labels. For the nodes without labels, we label them by 0. For each pair of nodes, x and y, every algorithm referred in this paper assigns a score as s_{xy}. For an unlabeled node u, the probability that it belongs to l_i is

$$p(l_i|u) = \frac{\sum_{\{v|v\neq u,\mathrm{label}(v)=l_i\}} s_{u,v}}{\sum_{\{v|v\neq u,\mathrm{label}(v)\neq 0\}} s_{u,v}}\,, \tag{13}$$

where $l_i \in L$. The predicted label of node u is determined by the largest $p(l_i|u)$. If there are more than one maximum values, we randomly select one. A simple example is shown in Fig. 1, where there are two kinds of labels (i.e. a and b) and five nodes, four of which are labeled already. Our task is to predict the label of the node 5. According to the common neighbors algorithm, we obtain the similarity between node 5 and the other four labeled nodes, and then we infer that the probability that node 5 is labeled by a equals 3/4.

To test the algorithm's accuracy, all the labeled nodes are randomly divided into two parts: the training set, V^T, is treated as known information, while the probe set, V^P, is used for testing. We denote q the proportion of labeled nodes divided into training set, which is considered as the density index. A smaller q indicates a sparser labeled network. The accuracy is quantified by the probability that we predict right. For a testing node $u \in V^P$ whose label is l_i, if $p(l_i) > p(l_j)$, $j \neq i$,

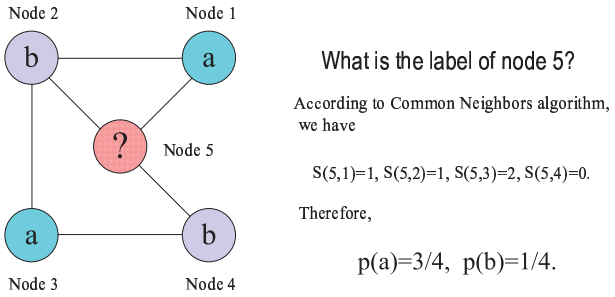


Fig. 1. (Color online) An illustration of how to predict the node's label according to the similarity.

we predict right, and thus $q_u = 1$. If there is n maximum values corresponding to n different labels and the right label is one of them, we have $q_u = 1/n$. Run over all the testing nodes and we have the accuracy equals

$$\text{Accuracy} = \frac{\sum_{u \in V^P} q_u}{|V^P|}, \tag{14}$$

where $|V^P|$ is the number of nodes in the probe set. For example, if there are two categories in the target network, namely l_1 and l_2, accuracy can be obtained by

$$\text{Accuracy} = \frac{n' + 0.5n''}{|V^P|}, \tag{15}$$

where n' is the number of nodes in probe set being predicted right and n'' is the number of nodes $u \in V^P$ having the same probability of two labels (i.e. $p(l_1|u) = p(l_2|u)$).

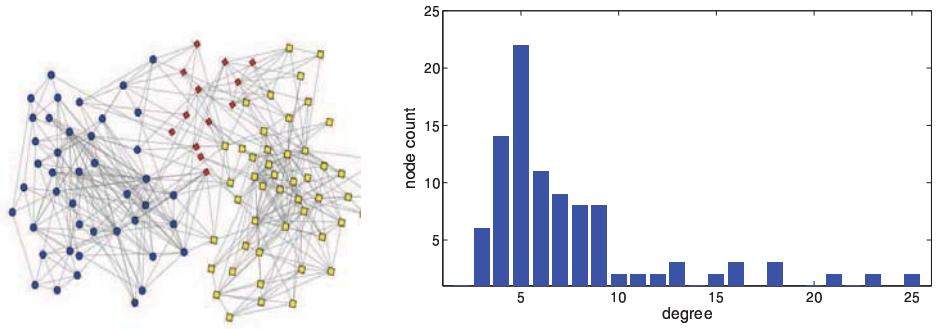
4. Empirical Results

We compare the above-mentioned ten similarity indices on the co-purchases network of political books.[19] This network contains 105 nodes (books) and 441 edges. All books are classified into three categories, neutral, liberal and conservative. The topological structure and the degree distribution of this network are shown in Figs. 2(a) and 2(b) respectively. Since the maximum degree is 25, there is no overwhelming hub in this network. The average shortest path and clustering coefficient of this network are 3.079 and 0.488 respectively. For simplicity, we start the experiments with the sampled networks containing only two classes. Therefore, we sample three labeled networks with three tasks as follows:

Task 1: *Whether an unlabel node is neutral*? For this task, we label the books which are neutral by a and others by b (i.e. not neutral).

Task 2: *Whether an unlabel node is liberal*? For this task, we label the books which are liberal by a and others by b (i.e. not liberal).

Task 3: *Whether an unlabel node is conservative*? We label the books which are conservative by a and others by b (i.e. not conservative).



(a) (b)

Fig. 2. (Color online) The topological structure of co-purchases network of political books (a) and its degree distribution (b). The red diamonds, blue circles and yellow squares stand for neutral, liberal and conservative nodes respectively.

Table 1. The summary of local consistency of each label and each sampled network. $N(a)$ and $N(b)$ are the number of nodes labeled by a and b respectively. $E(a)$ and $E(b)$ indicate the number of edges connecting to the nodes labeled by a and b respectively. $M(a)$ and $M(b)$ are the number of edges whose two endpoints have the same label a and b respectively. $C(a)$ and $C(b)$ are the local consistency of the nodes labeled by a and b respectively. C and C_2 are the local consistency and two-step consistency of the sampled network, respectively.

Net	$N(a)$	$N(b)$	$E(a)$	$E(b)$	$M(a)$	$M(b)$	$C(a)$	$C(b)$	C	C_2
Net1	13	92	67	432	9	374	0.134	0.866	0.869	0.864
Net2	43	62	208	269	172	233	0.827	0.866	0.918	0.894
Net3	49	56	236	251	190	205	0.805	0.817	0.890	0.882

Table 1 summarizes the basic statistics of these three sampled networks corresponding to tasks 1, 2 and 3, respectively. $N(x)$ $(x = a, b)$ is the number of nodes labeled by x. $E(x)$ indicates the number of edges connecting to the nodes labeled by x. Denote by $M(x)$ the number of edges whose two endpoints have the same label x, then $C(x) = M(x)/E(x)$ indicates the local consistency of the subgraph constituted by the nodes labeled by x and the edges connecting to these nodes. C is the local consistency of the whole network, which reads $C = (M(a) + M(b))/E$, where E is the total number of edges of the whole network (here $E = 441$). Note that, $E < E(a) + E(b)$. Here, we further develop the definition of local consistency to two-step consistency denoting by C_2 which equals to the number of paths with length 2 whose two endpoints have the same label divide by the number of the path with length 2. Clearly, the common neighbors index will perform well in the network with high C_2. Four simple examples of calculating $C(x)$, C and C_2 are shown in Fig. 3. One can see that in the first graph, because of $C = 0$, RN will perform very bad, while CN performs very good ($C_2 = 1$). However in the forth graph both RN and CN can give good performance.

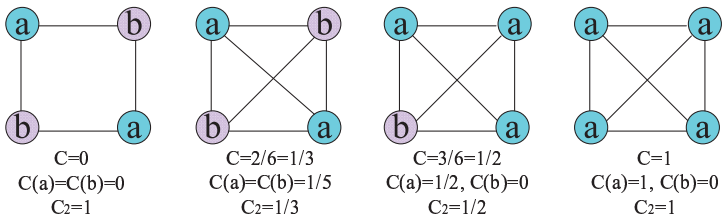


Fig. 3. (Color online) Illustration of the calculation of local consistency and two-step consistency.

Table 2. The dependence of local consistency C and two-step consistency C_2 on the proportion of training set q (range from 0.1 to 0.9).

Net $C(q)$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Net1	0.459	0.659	0.747	0.801	0.823	0.835	0.839	0.841	0.841
Net2	0.487	0.704	0.801	0.859	0.886	0.898	0.903	0.901	0.896
Net3	0.475	0.681	0.773	0.829	0.853	0.866	0.871	0.871	0.864
Net $C_2(q)$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Net1	0.705	0.797	0.818	0.829	0.833	0.835	0.835	0.837	0.839
Net2	0.731	0.826	0.851	0.863	0.866	0.868	0.869	0.867	0.869
Net3	0.712	0.806	0.831	0.843	0.848	0.852	0.853	0.855	0.856

Comparisons of the ten similarity indices on three sampled networks are shown in Fig. 4. The subgraphs (a), (c) and (e) show the results of the local indices, while (b), (d) and (f) report the results of the global indices. It is interesting that all these five local indices give almost the same results especially when the density of labeled nodes is small. This is because all these five indices are common-neighbor-based and when q is small whether an unlabeled node relevant with a labeled node plays a more important role than the exact correlation (similarity score) between them. Furthermore, all the common-neighbor-based indices perform well and even when the data is sparse they can give much better prediction than RN. This is because the local consistency C, which affects the performance of RN, is very sensitive to q, while the two-step consistency C_2, which affects the performance of CN, is not. The dependence of C and C_2 on the proportion of training set q is shown in Table 2, where one can find that when q changes from 0.2 to 0.1, C sharply decreases more than 30%, while C_2 decreases only 10%. In addition, in Fig. 4(c) RN performs better than CN when q is large. The reason is for Net2 C is much larger than C_2 when the data is dense (see Table 2).

Comparing with global indices, the best performing local index can give competitively accurate or sometimes even better classification when q is large. However, when the labeled data is sparse, for most unlabeled nodes it is too difficult to find a labeled node nearby, and thus the global indices will be potential to give better prediction. Actually when the data is sparse the best performing global index will predict better than the best performing local index. Among these five global

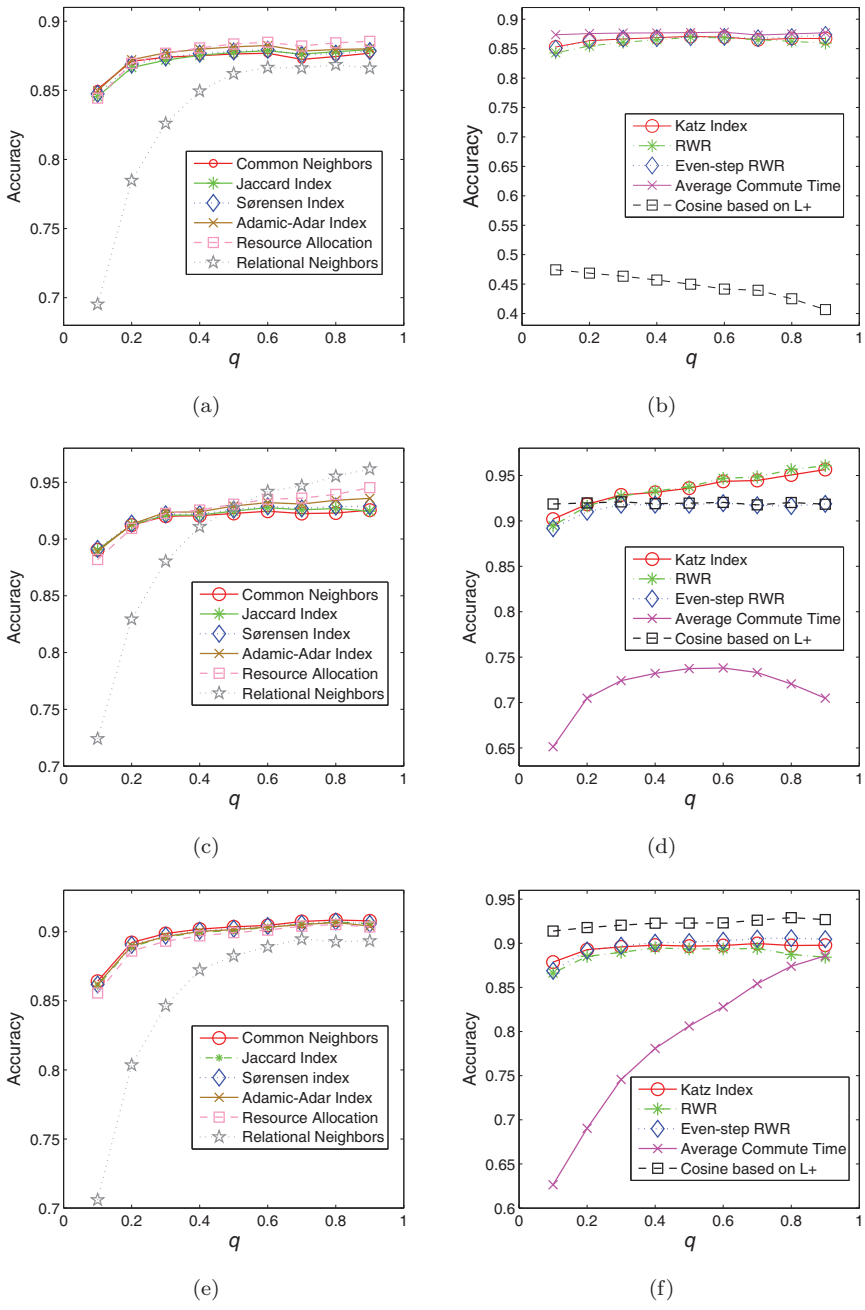


Fig. 4. (Color online) Comparison of ten similarity indices on three sampled networks containing two categories. (a) and (b) are the results of the local and global indices for task 1 respectively. (c) and (d) are the results of the local and global indices for task 2 respectively. (e) and (f) are the results of the local and global indices for task 3 respectively. For *RWR* index we set $c = 0.1$. Each number is obtained by averaging over 1000 implementations with independently random division of training set and probe set. The variance of calculation has a milli order of magnitude.

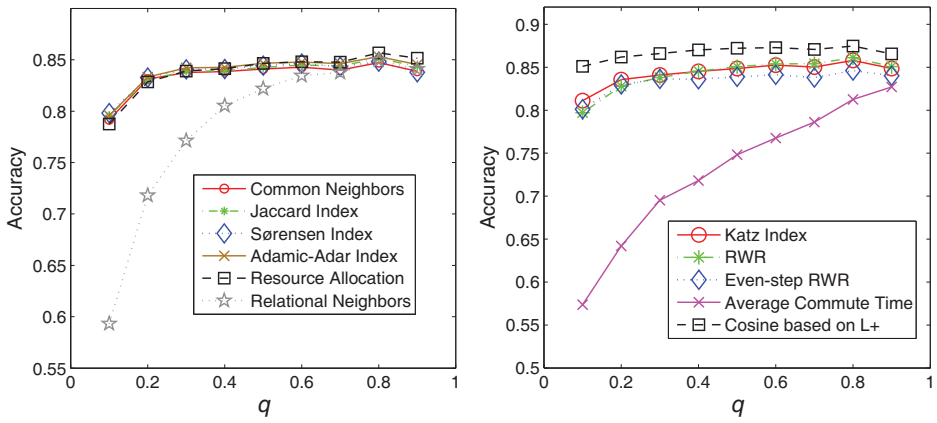


Fig. 5. (Color online) Comparison of ten similarity indices on the network taking into account three categories. For RWR we set $c = 0.1$. Each number is obtained by averaging over 1000 implementations with independently random division of training set and probe set.

indices, the performance of *Katz index*, *RWR* and *Even-step RWR* are stable, while the performance of ACT and $\cos^+$ are not. For example, in Net1, the ACT index performs very well but $\cos^+$ is even worse than pure chance. However, in Net3, the $\cos^+$ index preforms the best but the ACT index preforms even worse than the simplest method *RN*.

Obviously, it will be more difficult to obtain highly accurate classification when considering many categories together. We further carry out an experiment on the network containing all the three categories. Our task is to detect the category of an unlabeled book, namely is it neutral, liberal or conservative? We label the books by n (i.e. neutral), l (i.e. liberal) and c (i.e. conservative) according to their categories. The local consistency and two-step consistency of this network are 0.8413 and 0.8204 respectively, which are all lower than the three sampled networks containing only two classes, and thus the accuracy is also lower, as shown in Fig. 5. One can see that the results are similar to the one on the Net3 where the biggest class, conservative, is considered.

5. Conclusion and Discussion

In this paper, we investigated the similarity-based classification for partially labeled network. The basic assumption is that two nodes are more likely to have the same label if they are more similar to each other. We introduced ten similarity indices which have been widely used to solve the link prediction problem of complex networks, including five common-neighbor-based indices, namely *Common Neighbors* (CN), *Jaccard coefficient*, *Sørensen index*, *Adamic-Adar index* (AA) and *Resource Allocation index* (RA), and five global indices, namely *Katz index*, *Average Commute Time* (ACT), *cosine based on the Pseudoinverse of the Laplacian matrix* ($\cos^+$), *Random Walk with Restart* (RWR) and *Even-step RWR*. We carried out

the experiments on the co-purchase network of political books. The results showed that the similarity-based classification performs overall better than the *Relational Neighbors* algorithm, especially when the labeled nodes are sparse. Furthermore, we found that when the data is dense, the best-performing local index can predict as good as the global indices and sometimes even better. However when the data is sparse, the best-performing global index will predict much better than the local indices. Comparing with the former proposed algorithms, the group of similarity-based classification methods has three advantages: firstly, it can, to some extent, solve the sparse data problem; secondly, when the network consistency assumption is not held it can still give high accurate classification; thirdly, without any learning process this method has lower calculation complexity than other complicated methods.

However, there are still some open problems left. For example, what is the relation between the network structure (or the label structure) and the performance of each similarity index? In-depth analysis on the modeled networks may be helpful, where we can control the topological properties, the label density and the network consistency. Anyway, we hope this work can provide a novel view for the study of classification in partial labeled networks and we believe that there is still a large space for further contribution. For example, in order to avoid the superfluous information, one can only consider the top-k similar labeled nodes when calculating the probability. In addition, one can also use negative correlation in the adjacent matrix A directly, namely for the nonzero element in A if the nodes x and y have the different labels, we set $A_{xy} = -1$. To do this, we cannot only obtain the strength of the correlation between the unlabeled node and the labeled one but also know the correlation type, positive or negative.

Acknowledgments

We acknowledge Tao Zhou for his assistance of manuscript preparation. This work is partially supported by the Swiss National Science Foundation (200020-121848), the Sichuan Provincial Science and Technology Department (Grant No. 2010HH0002) and the National Natural Science Foundation of China (60973069, 90924011).

References

1. S. Chakrabarti, B. E. Dom and P. Indyk, in *Proc. SIGMOD-98, ACM International Conference on Management of Data* (ACM Press, Seattle, WA, 1998), p. 307.
2. Y. Yang, S. Slattery and R. Ghani, *J. Intell. Infor. Syst.* **18**, 219 (2002).
3. B. Gallagher, H. Tong, T. Eliassi-Rad and C. Faloutsos, in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM Press, New York, 2008), p. 256.
4. V. R. Carvalho and W. W. Cohen, in *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (ACM Press, New York, 2005), p. 345.
5. L. Zhang, K. Hu and Y. Tang, *Cent. Eur. J. Phys.*, DOI:10.2478/s11534-009-0114-9 (2009).

6. P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher and T. Eliassi-Rad, *AI Magazine* **29**, 93 (2008).
7. X. Zhu and A. B. Goldberg, *Synthesis Lect. Artif. Intell. Machine Learning* **3**, 1 (2009).
8. D. Liben-Nowell and J. Kleinberg, *J. Am. Soc. Inf. Sci. Technol.* **58**, 1019 (2007).
9. T. Zhou, L. Lü and Y.-C. Zhang, *Eur. Phys. J. B* **71**, 623 (2009).
10. L. Lü, C.-H. Jin and T. Zhou, *Phys. Rev. E* **80**, 046122 (2009).
11. F. Lorrain and H. C. White, *J. Math. Soc.* **1**, 49 (1971).
12. P. Jaccard, *Bull. Soc. Vaud. Sci. Nat.* **37**, 547 (1901).
13. T. Sørensen, *Biol. Skr.* **5**, 1 (1948).
14. L. A. Adamic and E. Adar, *Soc. Networks* **25**, 211 (2003).
15. L. Katz, *Psychmetrika* **18**, 39 (1953).
16. D. J. Klein and M. Randic, *J. Math. Chem.* **12**, 81 (1993).
17. S. Brin and L. Page, *Comput. Networks ISDN Syst.* **30**, 107 (1998).
18. S. Macskassy and F. Provost, in *Proc. 2nd International Workshop on Multi-Relational Data Mining* (ACM Press, New York, 2003), p. 64.
19. V. Krebs, *Int. Assoc. Human Resource Infor. Management J.* **4**, 87 (2000).